A Review: Various Hashing Algorithms in Data Deduplication for Storage Enhancement

Er. Karmjeet Singh[1], Er. Ramanjot Kaur[2]
[1]M.Tech (Scholar), [2]Assistant Professor
Doaba Institute of Engineering and Technology, Kharar (Punjab)

Abstract. Storage becomes most far-reaching when statics inject to cloud. By uploading diverse copies of facts having homogenous features and parameter from different sources over cloud-network, the recourse of monotonous statics enlarges at giant outlay. For a coherent processing speed of data in the cloud network, it's compulsory to reduce storage complexity and scrap. To accomplish this, statics are disintegration into trivial nabs. Computation is designed to identify the duplicate statics and remove them completely, but every program has some limitation to identify all parameter of monotonous facts... In this research paper, I am going to use updated and improved computation to identify the pattern of duplicate data for more accurate and efficient consequence as well as to enhance storage**.**

Keyword *- Data-duplication, monotonous statics, improved hashing algorithms and computation, hash function, deduplication optimized algorithm, data chunks, reduce data complexity.*

I. INTRODUCTION

Today and tomorrow are based on cloud computing. Every industry directly or indirectly always stays interconnected with computer and cloud storage. So the concept of cloud computing was introduced which plays a unique role in computing. Computing going on advancement every single second and huge amount of data uploaded to cloud continuously. Now here strenuous complication merges i.e. indexing of statics and searching relative records in a raw material efficiently. Data-Servers are overloaded to keep the record of each business transactions and everyday activity, business reports, office records, entertainment and memories etc… It's necessary to have larger and larger space for storage, but storage server has a particular limit which can't be exceeded. In every second multiple copies of facts uploaded to online cloud storage which is responsible to fulfill the requirement of daily life dealings. By doing so, data transfer rate and network congestion increase at very high tempo. Process of remote merging always continues, so hump of data-duplication, security issuance, slow data transfer rate, storage space complexity and management, have to face, because it's much difficult to index million billion trillion bytes of data in a perfect manner within short time period or seconds.

Fig. 1 Data Uploading in Cloud Network

To deal with these types of problems number of techniques introduced. Let's see first for a data security- encryption-decryption, plaintext, cipher text, fingerprint, public and private key concepts were used and to handle the problem of storage and data repetition, compression and hashing concepts were introduced. These techniques play the crucial role to optimize web services successfully.

But whenever any raw material uploaded to World Wide Web, it passes through the number of algorithms before to storing in the actual sever space. To tackle with this raw material, computations were designed, mostly helpful proved are hashing functions and chunks algorithms, where the smallest part of data called chunk and unorganized smallest part data called hash, a hash is small mapped form of data. Replication method technically broke up facts and figure in smaller piece i.e. chunks and hash computations to assign each chunk a hash value that is called a fingerprint, which was evaluated by already stored identifier to check the data already existed or not. If yes then informed back to computation, if not then fresh fingerprint is updated in database storage server. I am going to mix-up outputs of different hashing algorithms which will act as input for other algorithm (LZW, MD5, SHA, CDC, BCO,) to enhance processing time and storage throughput by compressing and reducing repetition of data.

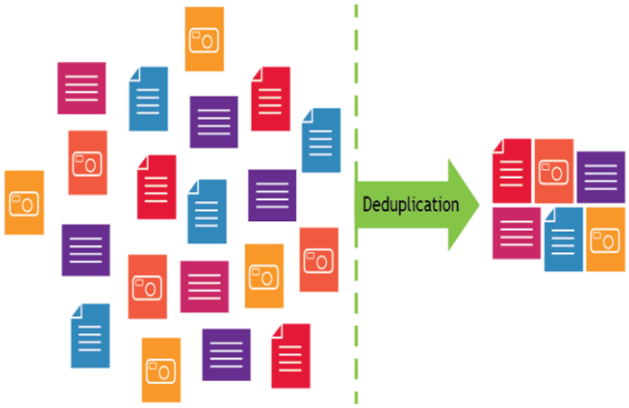

Fig. 2 Deduplication files optimization

II. LITERATURE

M. Armbrust et al. [1] Cloud web system is storage space where raw data are stored which are generally used by the third person. Cloud storage helps mobility of data anywhere it needed. It makes cloud storage cost effective as well as scalable service. These benefits of web storage attract more and more clients. According to recent analysis, the volume of statics will reach approx 45 trillion GB up to 2020. The main difference between web storage and manually server storage is that cloud transfer with the help domain whereas other directly to the local server and in online server repetition of files is detected by algorithms. Even web storage widely preferred it fails to manage uniqueness of data, Deduplication occurred while indexing files to the warehouse and also suffering from security threats from both internally and externally.

Haitao et al. [2] proposed relocation methods taking into account (dynamic, receptive & shrewd procedures), albeit basically in light of the present information, can make the combination cloud-helped VoDgroup set aside to 30% transmission capacity cost contrasted & the Clients/Server mode. They can likewise handle unpredicted the glimmer assembly activity with little cost. It likewise establishes that the cloud cost & server program capacity picked assume the essential parts in sparing expense, while the distributed storage size & cloud constituent upgrade system undertake the key parts in the client knowledge change.

Ward et al. [3] represented acquainted the augmentations with a coordinated mechanization volume called the Darwin construction that empowers on load undertaking for this situation & talks about the effect that computerized relocation has on the expense & dangers ordinarily connected with relocation to the cloud.

Kang et al. [4] proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capability. Then, if the relocation restriction is gratified, we transfer a different VM after this PM to oblige the novel VM. In addition, they are a mixed learning system where a lot is working to recognize forthcoming VMs for the on-line expansion. Assessment upshots establish greater competence of our method. TAR scheme also introduced in it.

Debnath BK et al. [5] duplication method works by broke down files into multiple chunks and assign value. This value is fingerprinting using 20 bytes of SHA-1 signature to check weather two value have identical data or not.

III. CHALLENGES IN DEDUPLICATION

- Research always continues to get most reliable solution as compared to previous obtained. Algorithms designed to process data when it uploads to cloud. A data can be easily processed when it's in lesser quantity. But becomes hardest problem when have to process million GB data within seconds. To process raw data online with high speed it's a biggest challenge to face in existing world.

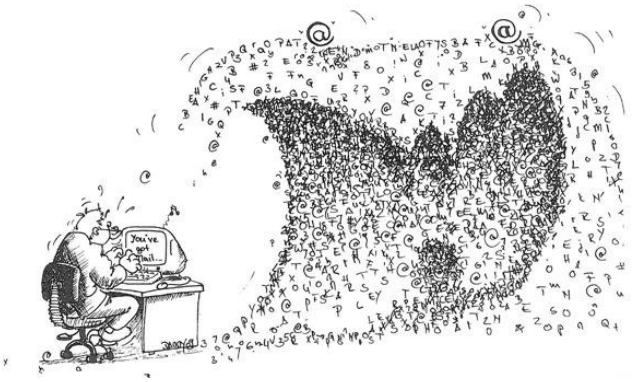

Fig.3 searching in billion records

- Now in a modern time period every business going online, and million – trillions TB data uploads to cloud storage. Is it actually easy to inject and fetch data from online server with high speed…? Can say somewhat…! Not at all, research is ongoing to get more optimized solution.
- Following are algorithms and computation which previous designed to reduce the d-duplication Secure has algorithm (SHA, SHA1, SH2, SH3), Cryptographic hash function, MD4, DSA, checksum algorithm (CRC), WEP encryption, Compression etc. but have some limitations.
- Due to high traffic search for desired result not obtained within time limits also difficult to do fragmentation quickly.
- Unwanted space consumption, and data is present in billion trillion.
- Some program good for searching in data, some are good in compression, some helps to deal with data bits, have to combine reach at optimal result.
- Current hashing function or searching technique is not much better. It is a more time consuming process to search any of the records or de-duplicate any new content.
- Loss of original data while storage and transferring due to corrupted bits.

- Data is the most important thing in the system so we need an accurate fingerprint generator algorithm which finds files fast and accurate and current system having this type of functions but isn't proposed 100% accuracy.

IV. ADVANTAGES IN DE-DUPLICATION

- Latest researches come with fruitful results to minimize the de duplication.
- Combination of algorithms helps to increase search efficiency, as well as reduce cost of system handling.
- Primary storage workloads (e.g. email, user directories, and databases) get benefits of de duplication, due to reduced latency cost.
- WAN (wide area network) bandwidth optimization done by reducing the number of bytes per second must be transferred between end points.
- Virtual severs also desktops got good sake from duplication because it allows separate system files for different virtual machine to be come together into a one storage space. Backing up and making duplicate files of virtual environments are also improved.
- Raw data compression can minimize number disk used in storage as well as for energy consumption costs. By analysis the data de-duplication strategy, processes, and implementations for further lay act as foundation of research work.

V. DISADVANTAGES IN DE-DUPLICATION

- Source and target differs in quantity and quality of data by removing duplicate copies.
- Duplicate copies sometime helps to maintain backup which also removed by de duplication. Whenever raw material is changed, problem arose about potential loss of data.
- There are number of methods for duplication with slightly different way. However, maximum chances to loss data integrity. While scaling disk space adversely affected.
- Data backup diminished while removing duplicates of copies of data. Primary and secondary storage affected.
- In case of de duplicity warehouse analyzed again and again which affect the efficiency and increase the cost of data maintenance as well as administrative time. And always worry about *Pay as you grow.*

VI. SEVERAL TECHNIQUES IN DE-DUPLICATION

A. Based on data de duplicated there are two methods in De-duplication

a) File Level De- duplication: - This method first detects the identical files and is removed. One copy of the file is stored. A Pointer is used to open the original file for the following copies. This method doesn't deliberate the contents present inside the file. For illustration, two manuscript files with simple title change are stored as two changed files. The benefit of this process is simple and fast. This method is also known as Single Instance Storage.

b) Block or Subfield De-duplication: -The File is divided into little chunks called blocks and duplicate blocks are detected using the specialized hash algorithm. If the data are unique written into disk else only pointer is used to point the disk location. According to the size of the block there are two processes in block De-duplication.

c) Fixed-length block: - De-duplication breaks the data into fixed size blocks. The disadvantage of this method is it fails to find the redundant data as a small change in the chunk result reworked of all subsequent blocks to the disk. But this method is fast, simple and minimum CPU overhead [7].

d) Variable-Length block: - De-duplication disruptions the data into adjustable size blocks. The improvement of this method is if any modification occurs the boundary of that block is improved and no change in consequent blocks and it saves more loading space when equated with fixed-length block De-duplication. This method requires more CPU cycles to identify block boundaries and for scanning entire file.

B. Based on implementations there are two methods in De-duplication

a) Source/Client based De-duplication: -The complete De-duplication process is done at source/client side before sending the data to a backup device. Only unique data are transferred to the backup device with the minimum available bandwidth and it needs less space.

b) Target based De-duplication:-The De-duplication process is done at back up device. When it receives the data with all its redundancy. This method needs more network bandwidth and it offloads backup client from deduplication process.

C. Constructed on when the De-duplication is done in back up device there are two methods

a) Inline De-duplication:- Allows De-duplication immediately after receiving the data at backup device. This method requires less storage capacity needed for backup.

b) Post-process De-duplication:- Allows De duplication after the received data is written into disk i.e. De-duplication is scheduled later. This method requires more storage space to store backup data.

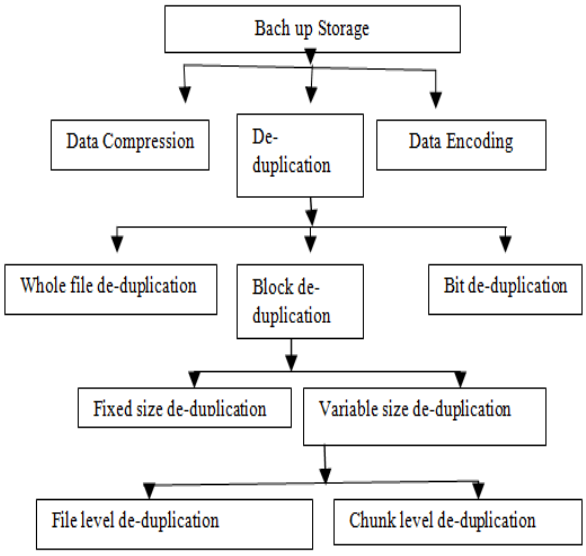

Fig. 4 Deduplication methods

Data deduplication describes a class of approaches that diminish the storage volume needed to store data or the amount of data that has to be transferred over a network. These approaches detect coarse-grained severances within a data set, mainly data de-duplication listening carefully in different terms like throughput, advance chunking schemes, the other type of storage capacity and cluster method and system workload.

1) SHA (SECUREHASH ALGORITHM)

Various types of SHA:-

SHA 0 SHA 1 SHA 256 SHA 512

The Protected Hash Algorithm is one of a number of cryptographic hash functions. There are presently three compeers of Secure Hash Algorithm [8].

SHA-1 is the original 160-bit hash function has a similarity to the earlier MD5 algorithm.

- SHA-2 is a relation of two similar hash functions, with different block sizes, known as SHA-256 and SHA-512. They differ in the word size; SHA-256 customs 32-bit words where SHA-512 usages 64-bit words.
- SHA-3 is a coming hash purpose standard still in development.

*2) Algorithm*1- the SHA algorithm uses 5 state variables, each of which is a 32 bit number (an unidentified long on most systems). These variables are shared and dice and are (eventually) the message digest. The variables are initialized as follows:

h0 = 0x67452301 h1 = 0xEFCDAB89

h2 = 0x98BADCFE h3 = 0x10325476

h4 = 0xC3D2E1F0 There are 80 rounds in SHA Algorithm.

VII. CONCLUSION AND FUTURE SCOPE

In this unique paper, we have merged algorithms and computation of D-duplication to get desired result with higher efficiency and fewer time, cost. The initial stage duplication split up raw data files for ease of processing is called chunks and further hash. Where fixed and variable sized chunking method ceases to work properly, we used improved SHA series algorithms and hashing techniques to overcome the throughput. This approach is made based on de duplication hash function which helps provide good throughput as well as decrease the space on occupied by repetitive file on cloud server. It worked fruitfully while uploading file on cloud server. This approach can be applied for modern generation devices like smart phone and tab as well for cloud storage. It can apply for heavy multimedia data like video, audio, image files etc. it is much simple to understand and it's throughput can be merged with another computation to get better results in future. It is an independent, flexible and cost-effective. It can be tested on encrypted and compressed where deduplication fails.

VIII. REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," Communication of the ACM, vol. 53, no. 4, pp. 50–58, 2010.

[2] Haitao Li, LiliZhong, Jiangchuan Li, , Bo Li, KeXu, " Cost-effective Partial Migration of VoD Services toContent Clouds", 2011 IEEE 4th International Conference on Cloud Computing, pp. 203-110, 2011.

[3] C. Ward, N. Aravamudan, K. Bhattacharya, K. Cheng, R. Filepp, R. Kearney, B. Peterson, L. Shwartz, C. C. Young, "Workload Migration into Clouds – Challenges, Experiences, Opportunities", 2010 IEEE 3rd International Conference on Cloud Computing, pp. 164-171, 2010.

[4] Kangkang Li, HuanyangZheng, &JieWu . "Migration-based Virtual Machine Placement in Cloud Systems", 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE, pp. 83-90, 2013.

[5] Debnath, B.K., Sengupta, S. and Li, J., 2010, June. ChunkStash: Speeding Up Inline Storage Deduplication Using Flash Memory. In USENIX annual technical conference (pp. 1-16).

[6] Gauravaram, P., Knudsen, L.R., Matusiewicz, K., Mendel, F., Rechberger, C., Schläffer, M. and Thomsen, S.S., 2009.

[7] Deshmukh, V. and Mu, P.Y., NetApp Inc, 2012. System and method for providing variable length deduplication on a fixed block file system.

[8] Gauravaram, P., Knudsen, L.R., Matusiewicz, K., Mendel, F., Rechberger, C., Schläffer, M. and Thomsen, S.S., 2009. Grøstl-a SHA-3 candidate.